ELMER PRESS

Check for updates

Original Article World J Oncol. 2023;14(5):406-422

Development of a Machine Learning-Based Prognostic Model for Hormone Receptor-Positive Breast Cancer Using Nine-Gene Expression Signature

Takashi Takeshita[a], Hirotaka Iwase[a], Rongrong Wu[b], Danya Ziazadeh[b], Li Yan[c], Kazuaki Takabe[b, d, e, f, g, i]

Abstract

Background: Determining the prognosis of hormone receptor positive (HR$^+$) breast cancer (BC), which accounts for 80% of all BCs, is critical in improving survival outcomes. Stratifying individuals at high risk of BC-related mortality and improving prognosis has been the focus of research for over a decade. However, these tools are not universal as they are limited to clinical factors. We hypothesized that a new framework for predicting prognosis in HR$^+$ BC patients can develop using artificial intelligence.

Methods: A total of 2,338 HR$^+$ human epidermal growth factor receptor 2 negative (HER2$^-$) BC cases were analyzed from Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), The Cancer Genome Atlas (TCGA), and Gene Expression Omnibus (GEO) cohorts. Groups were then divided into high- and low-risk categories utilizing a recurrence prediction model (RPM). An RPM was created by extracting nine prognosis-related genes from over 18,000 genes using a logistic progression model.

Results: Risk classification by RPM was significantly stratified in both the discovery cohort and validation cohort. In the time-dependent area under the curve analysis, there was some variation depending on the cohort, but accuracy was found to decline significantly after about 10 years. Cell cycle related gene sets, MYC, and PI3K-AKT-mTOR signaling were enriched in high-risk tumors by the Gene Set Enrichment Analysis. High-risk tumors were associated with high levels of immune cells from the lymphoid and myeloid lineage and immune cytolytic activity, as well as low levels of stem cells and stromal cells. High-risk tumors were also associated with poor therapeutic effects of chemotherapy and endocrine therapy.

Conclusions: This model was able to stratify prognosis in multiple cohorts. This is because the model reflects major BC therapeutic target pathways and tumor immune microenvironment and, further is supported by the therapeutic effect of chemotherapy and endocrine therapy.

Keywords: Breast cancer; Recurrence prediction; Cancer genomics; Tumor immune microenvironment; Machine learning

Introduction

Breast cancer (BC) is currently one of the most common types of cancer in women, and is increasing in incidence by 0.5% per year [1]. Approaches to improve the prognosis of hormone receptor positive (HR$^+$) BC, which accounts for 80% of all BCs, is vital to enhancing BC survival. Improving stratification of individuals at high risk of BC-related mortality has been the focus of research interest for over a decade [2]. Researchers have mainly classified BC patients by different gene expression profiles and stratified according to clinical outcome [3, 4]. Such studies have evolved into the development of algorithms for estimating the risk of recurrence and survival by molecular gene expression signatures [5]. Some of these molecular prognostic indicators have since been recommended by the guidelines of ASCO and are now available for the clinical management of BC.

Another approach to improve the outcome of HR+ BC is identification of predictive biomarkers. For HR+ human epidermal growth factor receptor 2 negative (HER2$^-$) early-stage BC, endocrine therapy (ET) is the main treatment, and the indication for adjuvant chemotherapy (CT), which has serious side effects, should be determined on a patient-by-patient basis. Therefore, additional prognostic information is often needed to provide patients with reliable and effective treatment. Five prognostic signatures

Manuscript submitted August 4, 2023, accepted August 28, 2023
Published online September 20, 2023

[a]Department of Breast and Endocrine Surgery, Kumamoto City Hospital, Kumamoto, Japan
[b]Breast Surgery, Department of Surgical Oncology, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA
[c]Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA
[d]Department of Surgery, University at Buffalo Jacobs School of Medicine and Biomedical Sciences, the State University of New York, Buffalo, NY, USA
[e]Department of Breast Surgery and Oncology, Tokyo Medical University, Tokyo, Japan
[f]Department of Surgery, Yokohama City University, Yokohama, Japan
[g]Department of Surgery, Niigata University Graduate School of Medical and Dental Sciences, Niigata, Japan
[h]Department of Breast Surgery, Fukushima Medical University, Fukushima, Japan
[i]Corresponding Author: Kazuaki Takabe, Breast Surgery, Department of Surgical Oncology, Roswell Park Comprehensive Cancer Center, Buffalo, NY 14263, USA. Email: kazuaki.takabe@roswellpark.org

doi: https://doi.org/10.14740/wjon1700

Table 1. Key resources

Resource	Source	Identifier
Deposited data		
METABRIC	METABRIC	[31]
TCGA	TCGA PanCancer Atlas	[31]
GSE199135	Takeshita et al [24]	[32]
GSE9195; GSE6532	Loi et al, 2010 dataset [25]	[32]
GSE21653	Sabatier et al, 2011 [26]	[32]
Software and algorithms		
Python 3.11.0	Python Software Foundation	[33]
Numpy v 1.23.4	Van Der Waltetal, 2011 [27]	[34]
SciPy v 1.9.3	Virtanen et al, 2020 [28]	[35]
Pandas v 1.5.1	Pandas - Python Data Analysis Library	[36]
Seaborn v 0.12.1	Waskom, 2021 [29]	[37]
Matplotlib v 3.6.2	Hunter, 2007 [30]	[38]
R4.0.2	The R Foundation	[39]

for BC (OncotypeDX®, MammaPrint®, Prosigna®, EndoPredict®, and Breast Cancer Index^{SM}) which are included in national and international guidelines (NCCN, ASCO, ESMO, NICE, AGO, and St. Gallen) are representative and have been summarized in a recent review [5]. In particular, Oncotype DX® has been successful in focusing on the additional benefits of CT for HR+ BC [6-8].

However, these tools are not universal, as they are limited to clinical factors such as HRs, menopausal status, and nodular status [9]. It remains unclear which gene assay should be prioritized, and prognosis and prediction differ between tests [5]. There remains a significant need for an unbiased and comprehensive approach to identify and list all prognosis-related molecules [5, 10-12]. Innovations in high-throughput technology have led to the rapid accumulation of data on gene expression throughout the transcriptome of tumors from large numbers of patients. The study of such accumulated data by the research community is a very important study to identify more accurate potential biomarkers at the individual patient level [13]. With these cohorts and algorithms, we have conducted studies that assess the real-world relevance of expression of genes of interest [14-24].

We aimed to test the hypothesis that artificial intelligence can be used to develop a new framework for predicting prognosis in BC patients in multiple validation cohorts.

Materials and Methods

In all cohorts, given that the patient data are de-identified, and that it is in a public domain, it waived Institutional Review Board approval.

Key resources are shown in Table 1 [24-39].

Study design and cohorts

We performed a retrospective analysis of six independent HR^+ $HER2^-$ BC cohorts, which included 1,355 women from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort, 585 women from The Cancer Genome Atlas (TCGA) cohort, and a total of 398 women from the Gene Expression Omnibus (GEO) cohorts (GSE199135, GSE9195, GSE6532, and GSE21653) in which transcriptomic data were associated with clinical parameters, all of which were previously published [14-24, 40]. We downloaded public data from cBioPortal [41, 42] with the CGDS-R package for METABRIC (METABRIC Nature 2012 and Nat Commun 2016 dataset) and TCGA (TCGA PanCancer Atlas dataset) and the GEOquery package for GEO cohorts. We added clinical data from Rueda et al, 2019 dataset for METABRIC [43], Liu et al, 2018 dataset for TCGA [44], Takeshita et al, 2022 dataset for GSE199135 [24], Loi et al, 2010 dataset for GSE9195 and GSE6532 [25], and Sabatier et al, 2011 dataset for GSE [26] to each cohort (Supplementary Material 1, www.wjon.org).

The initial analysis was conducted with the METABRIC cohort (discovery cohort), given that this is the best-characterized cohort available. We then performed gene analysis differentially expressed with |log2 fold-change| greater than 0.25 and adjusted $P < 0.05$ in distant recurrence and identified 155 genes (Supplementary Material 2, www.wjon.org). For the 155 extracted human protein-coding genes, we used the Cox proportional-hazards (Cox-PH) model to examine the potential utility of each gene as a prognostic marker. Prognostic-related genes were defined as genes with P-values less than 0.05 in the Cox-PH model. Consequently, we uncovered 23 genes closely involved in distant recurrence free survival (RFS) in METABRIC HR^+ $HER2^-$ patients (Supplementary Material 3, www.wjon.org). Furthermore, we combined the Cox-PH model with recursive feature elimination to narrow down the nine best predictors (Table 2). After validating several artificial intelligence-based machine learning algorithms, we used a logistic regression model (LRM) to optimize the weights of the nine selected genes and build an RPM. This model was generated using the Python-based scikit-learn library.

Table 2. The Nine Genes Best Predictors Extracted From 23 Signature Genes Using the Cox-PH Model With Recursive Feature Elimination

	coef	std err	z	P > |z|	(0.025	0.975)
23 genes						
const	-0.568	0.684	-0.83	0.406	-1.909	0.773
AGL	0.2822	0.109	2.591	0.01	0.069	0.496
BIRC5	0.1692	0.227	0.747	0.455	-0.275	0.613
C1orf64	-0.216	0.061	-3.565	0	-0.335	-0.097
CDCA3	0.1451	0.282	0.514	0.607	-0.408	0.698
CENPF	0.1927	0.243	0.795	0.427	-0.283	0.668
CEP55	-0.7396	0.378	-1.957	0.05	-1.48	0.001
CIDEC	0.021	0.077	0.272	0.785	-0.13	0.172
CKAP2L	0.8004	0.44	1.82	0.069	-0.062	1.663
CRTAP	-0.3557	0.197	-1.808	0.071	-0.741	0.03
CYP4F22	-0.1655	0.09	-1.845	0.065	-0.341	0.01
E2F2	-0.3632	0.253	-1.434	0.152	-0.86	0.133
FHL2	-0.0056	0.098	-0.057	0.955	-0.197	0.186
FOS	0.0434	0.077	0.566	0.571	-0.107	0.194
GSTM2	-0.0754	0.08	-0.937	0.349	-0.233	0.082
HNMT	-0.5129	0.213	-2.405	0.016	-0.931	-0.095
KIF20A	1.2829	0.324	3.956	0	0.647	1.919
LAD1	0.2009	0.083	2.423	0.015	0.038	0.363
PIP	0.0504	0.044	1.155	0.248	-0.035	0.136
PRC1	-0.5401	0.27	-2	0.045	-1.069	-0.011
S100P	0.1684	0.047	3.571	0	0.076	0.261
SEPP1	0.2943	0.13	2.267	0.023	0.04	0.549
STAT1	-0.0774	0.111	-0.695	0.487	-0.296	0.141
TUBA3D	-0.2673	0.074	-3.619	0	-0.412	-0.123
13 genes						
const	-0.93	0.458	-2.029	0.042	-1.828	-0.032
AGL	0.2695	0.105	2.557	0.011	0.063	0.476
C1orf64	-0.2189	0.058	-3.774	0	-0.333	-0.105
CEP55	-0.7108	0.353	-2.012	0.044	-1.403	-0.019
CKAP2L	0.7766	0.4	1.943	0.052	-0.007	1.56
CRTAP	-0.345	0.184	-1.876	0.061	-0.705	0.015
CYP4F22	-0.1731	0.088	-1.96	0.05	-0.346	-3.02E-05
HNMT	-0.3755	0.198	-1.893	0.058	-0.764	0.013
KIF20A	1.2816	0.307	4.168	0	0.679	1.884
LAD1	0.2107	0.081	2.611	0.009	0.053	0.369
PRC1	-0.513	0.261	-1.966	0.049	-1.024	-0.002
S100P	0.1733	0.046	3.74	0	0.082	0.264
SEPP1	0.2812	0.123	2.277	0.023	0.039	0.523
TUBA3D	-0.2489	0.07	-3.532	0	-0.387	-0.111
12 genes						
const	-0.9519	0.458	-2.079	0.038	-1.85	-0.054
AGL	0.3102	0.103	3.013	0.003	0.108	0.512
C1orf64	-0.2124	0.058	-3.668	0	-0.326	-0.099
CEP55	-0.631	0.35	-1.803	0.071	-1.317	0.055

Table 2. The Nine Genes Best Predictors Extracted From 23 Signature Genes Using the Cox-PH Model With Recursive Feature Elimination - *(continued)*

	coef	std err	z	P > |z|	(0.025	0.975)
CKAP2L	0.9341	0.392	2.383	0.017	0.166	1.702
CYP4F22	-0.17	0.088	-1.926	0.054	-0.343	0.003
HNMT	-0.5147	0.184	-2.793	0.005	-0.876	-0.153
KIF20A	1.1914	0.303	3.931	0	0.597	1.785
LAD1	0.2105	0.081	2.603	0.009	0.052	0.369
PRC1	-0.5702	0.259	-2.202	0.028	-1.078	-0.063
S100P	0.1651	0.046	3.582	0	0.075	0.255
SEPP1	0.183	0.112	1.636	0.102	-0.036	0.402
TUBA3D	-0.2344	0.07	-3.353	0.001	-0.371	-0.097
11 genes						
const	-0.8428	0.452	-1.866	0.062	-1.728	0.042
AGL	0.3101	0.103	3.014	0.003	0.108	0.512
C1orf64	-0.2005	0.057	-3.495	0	-0.313	-0.088
CEP55	-0.5666	0.347	-1.632	0.103	-1.247	0.114
CKAP2L	0.8471	0.388	2.182	0.029	0.086	1.608
CYP4F22	-0.1772	0.088	-2.014	0.044	-0.35	-0.005
HNMT	-0.3585	0.157	-2.289	0.022	-0.665	-0.051
KIF20A	1.1552	0.302	3.822	0	0.563	1.748
LAD1	0.2013	0.08	2.501	0.012	0.044	0.359
PRC1	-0.5527	0.259	-2.134	0.033	-1.06	-0.045
S100P	0.1609	0.046	3.504	0	0.071	0.251
TUBA3D	-0.2376	0.07	-3.398	0.001	-0.375	-0.101
10 genes						
const	-0.7569	0.447	-1.692	0.091	-1.633	0.12
AGL	0.2979	0.102	2.907	0.004	0.097	0.499
C1orf64	-0.1867	0.057	-3.294	0.001	-0.298	-0.076
CKAP2L	0.5869	0.352	1.67	0.095	-0.102	1.276
CYP4F22	-0.1749	0.088	-1.99	0.047	-0.347	-0.003
HNMT	-0.4006	0.154	-2.602	0.009	-0.702	-0.099
KIF20A	1.036	0.292	3.545	0	0.463	1.609
LAD1	0.1867	0.08	2.341	0.019	0.03	0.343
PRC1	-0.6683	0.249	-2.688	0.007	-1.156	-0.181
S100P	0.1609	0.046	3.509	0	0.071	0.251
TUBA3D	-0.2381	0.07	-3.416	0.001	-0.375	-0.101
9 genes						
const	-1.0651	0.407	-2.617	0.009	-1.863	-0.267
AGL	0.3107	0.102	3.04	0.002	0.11	0.511
C1orf64	-0.1932	0.057	-3.416	0.001	-0.304	-0.082
CYP4F22	-0.176	0.088	-2.009	0.045	-0.348	-0.004
HNMT	-0.4024	0.154	-2.611	0.009	-0.705	-0.1
KIF20A	1.2687	0.258	4.915	0	0.763	1.775
LAD1	0.1841	0.08	2.309	0.021	0.028	0.34
PRC1	-0.4968	0.226	-2.2	0.028	-0.939	-0.054
S100P	0.1645	0.046	3.596	0	0.075	0.254
TUBA3D	-0.2352	0.07	-3.38	0.001	-0.372	-0.099

Cox-PH: Cox proportional-hazards.

To explore whether this model can stratify prognosis, we analyzed the METABRIC total recurrence, local recurrence cohort, TCGA BC cohort, and another independent BC cohort, GSE199135, GSE9195, GSE6532, and GSE21653 to verify its performance.

Cluster analysis

We chose hierarchical clustering using the Euclidean distance and Ward's linkage due to its relative good performance [45]. The R-function "hclust" was used for performing hierarchical clustering.

Gene Set Enrichment Analysis (GSEA)

GSEA was performed comparing high and low risk of recurrence in RPM among hallmark gene sets using software provided by the Broad Institute [46], as we described previously [14, 15, 17]. We only considered gene sets significantly enriched that met a threshold of normalized enrichment score (NES) > 1.6 or < -1.6 and false discovery rate (FDR) q-value < 0.025. Gene Set Variation Analysis (GSVA) from the MSigDB Hallmark collection [47] was used to score the cancer hallmark gene sets for analysis. In doing so, we used the GSVA Bioconductor package (version 3.10), as we described previously [24].

Tumor microenvironment (TME) analysis

xCell, which is the bioinformatics tool that performs cell type enrichment analysis from gene expression data for 64 immune and stroma cell types, was used for TME analysis [48], as we described previously [24]. The immune cytolytic activity was defined as the geometric mean of GZMA and PRF1 expression values in Transcripts Per Million [49, 50] and immune cytolytic activity was calculated as previously described [16-24].

Statistical analysis

All statistical analyses were performed using R software [39] and Bioconductor [51] and Python (version 3.10.7 [33]). The Chi-square test or Fisher's exact test or the nonparametric Mann-Whitney U test and contingency analysis were used to assess baseline differences between binary variables. In the analysis of RFS, the Kaplan-Meier method was used to estimate survival rates, and differences between survival curves were evaluated by the log-rank test. Two-sided P-value < 0.05 was considered as statistically significant for all tests.

Results

Extraction of all prognosis-related genes and preparation of an RPM using machine learning

In order to build an RPM, we examined the relation between abundance of mRNA expression and distant recurrence with METABRIC as a discovery cohort. Figure 1a shows volcano plots that represent the distribution of the fold changes and adjusted P-values of 18,484 genes in METABRIC HR$^+$ HER2$^-$ cohort with and without distant recurrence. We identified 155 distant recurrence-related genes, which were differentially expressed with |log2 fold-change| greater than 0.25 and adjusted P < 0.05, with the complete list of these genes and their adjusted P values being provided in Supplementary Material 2 (www.wjon.org). A bidirectional hierarchical clustering heatmap, based on the expression levels of the identified differentially expressed genes (DEGs), indicated the relationship between clustering of the samples into two groups and type of recurrence (Fig. 1b). Analysis of the combined hazard ratio logarithms of all 155 extracted genes suggests that further refinement is necessary for prognostic prediction (Fig. 1c). Using the Cox-PH model, the optimal combination of prognostic genes was further screened from the 155 feature genes. Consequently, 23 DEGs were revealed to be closely involved in METABRIC HR$^+$ HER2$^-$ patients' distant RFS (P<0.05) (Supplementary Material 3, www.wjon.org). Combining the Cox-PH model and recursive feature elimination, we sequentially removed genes with P ≥ 0.05 from the 23 genes, and finally extracted the nine best predictors with P < 0.05 (Table 2).

In these nine best predictors, KIF20A and PRC1 were the most promising prognosis-related genes, with the highest and lowest hazard ratios, respectively. It is of note that one of these two genes, KIF20A, was not adopted by the MammaPrint or Oncotype Dx 21-gene RS systems. Indeed, most of the validated prognosis-related genes have not been well characterized to date with regard to their relation to basic or clinical oncology, with CYP4F22, TUBA3D, and HNMT receiving even less study.

Further, we investigated whether these nine newly identified prognosis-related genes were sufficient to predict survival in BC patients. An LRM was used to optimize the weights of the nine selected genes to build an RPM. In the METABRIC HR$^+$ HER2$^-$ cohort as the training cohort, the model was set up to significantly discriminate the high risk of recurrence group with the log-rank P value of < 0.00001 for the Kaplan-Meier survival curve (Fig. 1d).

Validation of RPM

To validate whether this RPM can universally stratify prognosis, we examined TCGA HR$^+$ HER2$^-$ BC cohort and other independent HR$^+$ HER2$^-$ BC cohorts, GSE199135, GSE9195, GSE6532, and GSE21653, using the Kaplan-Meier method and verified by the log-rank test. The Kaplan-Meier survival curve shows that the log-rank P-values for RFS time in the validation sets were all < 0.05 (Fig. 2), suggesting significantly different RFS time between predicted recurrence and non-recurrence samples. A time-dependent area under the receiver operating characteristics curve (AUC) value was created to examine the details of each cohort's timely accuracy (Supplementary Material 4, www.wjon.org). The AUC values peaked above 0.8 at GSE199135, GSE6532, and GSE21653. GSE199135 and GSE6532 main-

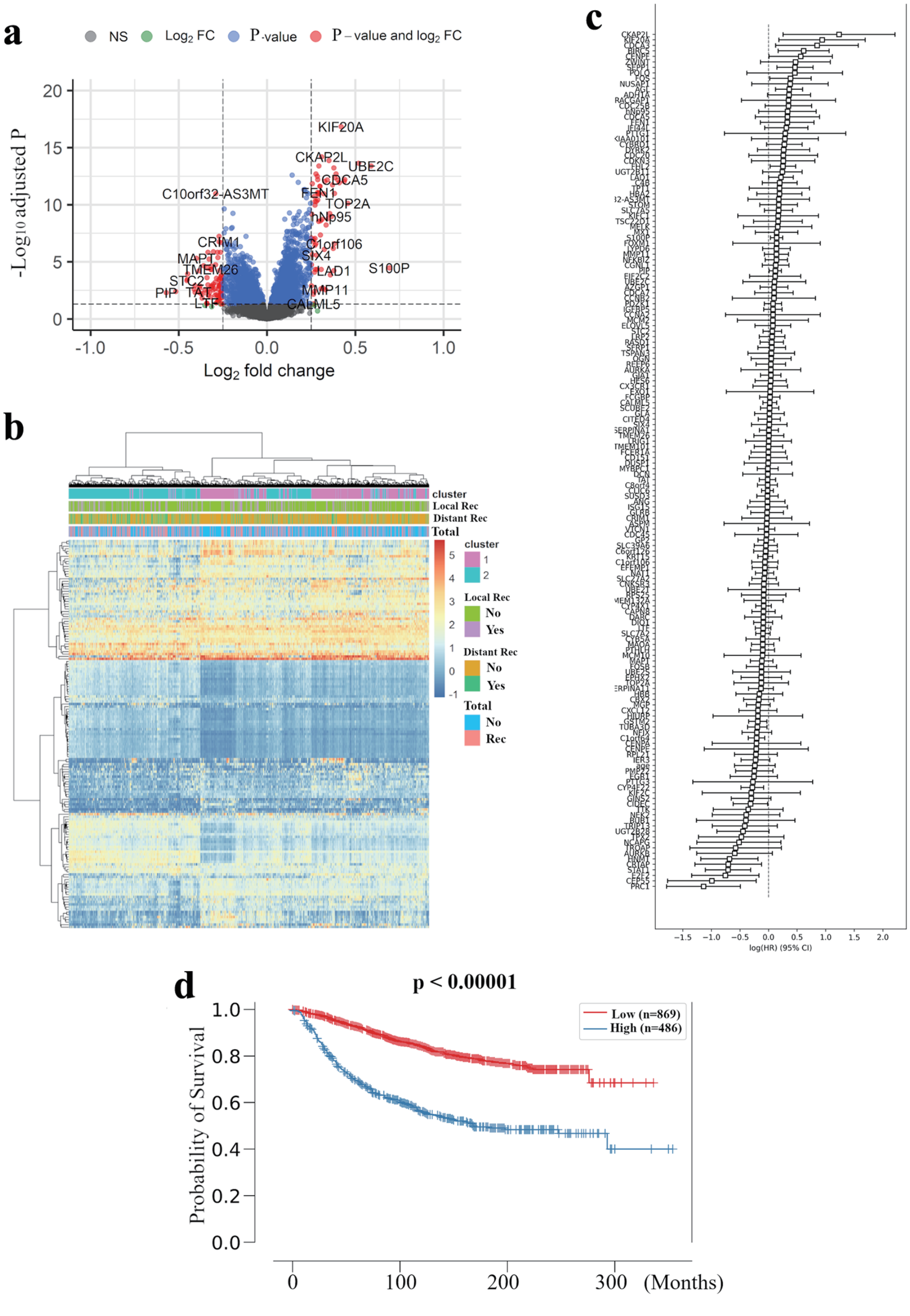

Figure 1. Extraction of all prognosis-related genes and preparation of a recurrence prediction model using machine learning. (a) Volcano plot illustrating the differentially expressed mRNAs of BC patients comparing with and without distant recurrence in the METABRIC HR$^+$ HER2$^-$ cohort are shown. X-axes: log2 FC; Y-axes: -log 10 adjusted P-value from limma analysis. mRNAs with adjusted P-value < 0.05 and log2 FC > 0.25 are marked in red, with adjusted P-value > 0.05 and log2 FC < 0.25 in green, with adjusted P-value < 0.05 and log2 FC < 0.25 in blue, all others in black. (b) A heatmap illustrating the expression intensity of 155 genes extracted by (a), with colors ranging from red to blue as indicated in the key are shown. Both rows and columns are clustered using correlation distance and average linkage. (c) Logarithm of the integrated hazard ratio for all 155 genes extracted by (a) are shown. The complete list of these genes identified by meta-analysis is provided in Supplementary Material 1 (www.wjon.org). (d) Kaplan-Meier curves for distant RFS in METABRIC HR$^+$ HER2$^-$ patients based on high and low risk in recurrence prediction model are shown. BC: breast cancer; METABRIC: Molecular Taxonomy of Breast Cancer International Consortium; FC: fold change; RFS: recurrence free survival; HR$^+$: hormone receptor positive; HER2: human epidermal growth receptor 2; LRM: logistic regression model.

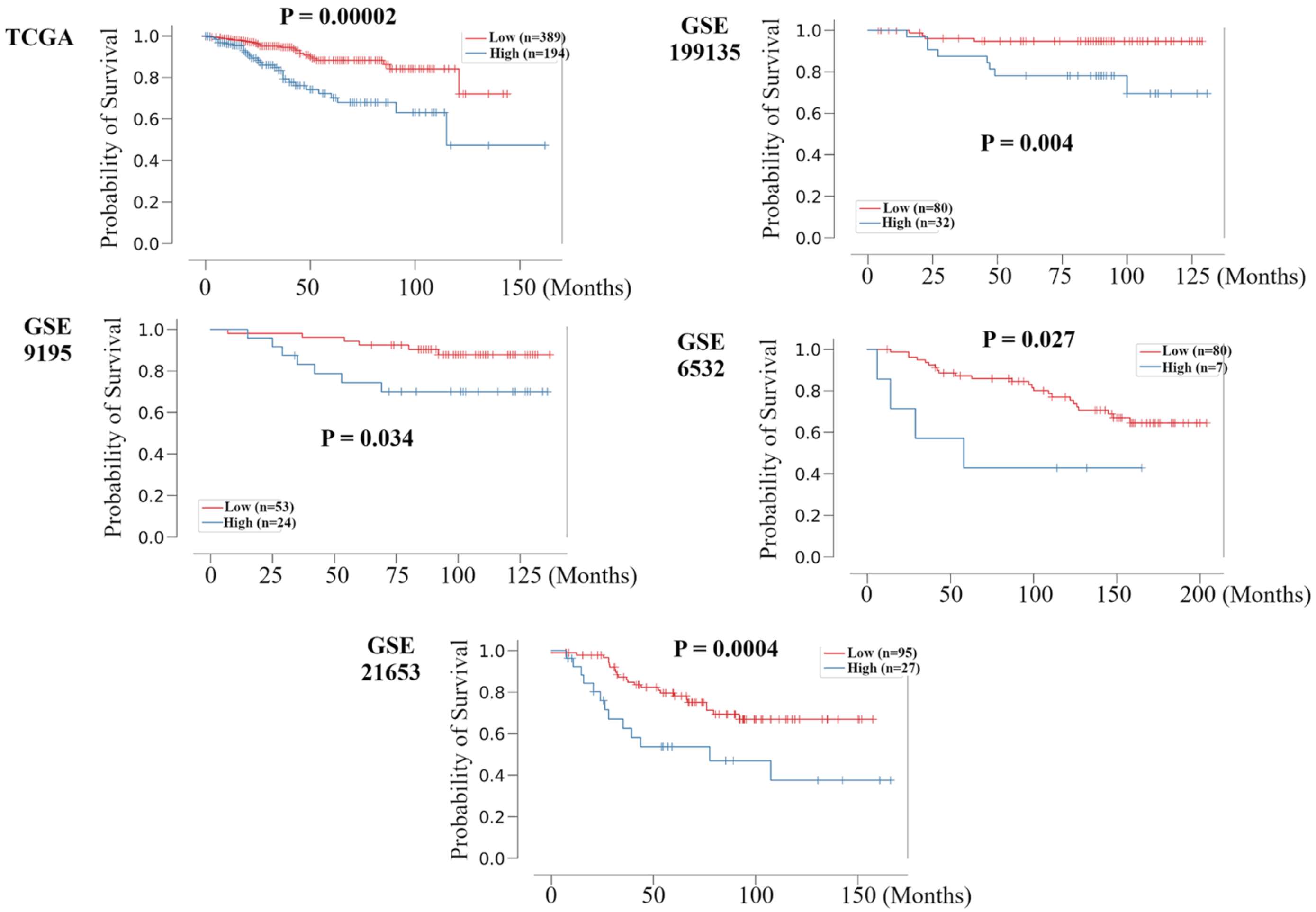

Figure 2. Validation of the relationship between the recurrence prediction model and survival rate in other HR$^+$ HER2$^-$ BC cohorts. Kaplan-Meier plots of the association of the recurrence prediction model with RFS for the recurrence prediction model, applied on the TCGA, GSE199135, GSE9195, GSE6532, and GSE21653 are shown. HR$^+$: hormone receptor positive; HER2: human epidermal growth receptor 2; BC: breast cancer; RFS: recurrence free survival; TCGA: The Cancer Genome Atlas; FC: fold change.

tained accuracy for about 10 years, but GSE21653 decreased in accuracy after 5 years. In METABRIC, TCGA, and GSE9195, the peak had not exceeded 0.8, but accuracy was also maintained for about 10 years. These results suggest that the RPM can universally and predominantly stratify prognosis in all HR$^+$ HER2$^-$ BC cohorts. In the time-dependent AUC analysis, there was some variation depending on the cohort, but accuracy tended to decline significantly after about 10 years.

High-risk tumors were associated with more aggressive clinical phenotype

Next, we used various factors to justify the risk classification by RPM as follows.

We studied the relationship between clinical features of the primary tumor and high and low risk in RPM amongst the HR$^+$ HER2$^-$ subgroup in the METABRIC cohort (Table 3). Patients categorized as high risk in the RPM were significantly associated with increased age, postmenopausal state, higher tumor size, lymph node metastasis, invasive ductal carcinoma, higher grade, progesterone receptor (PgR) positivity, luminal B subtype, HER2 subtype, and basal-like subtype.

Cell cycle related gene sets, MYC, and PI3K-AKT-mammalian target of rapamycin (mTOR) signaling were enriched in high-risk tumors in the GSEA

In order to investigate which mechanism is related to the RPM, we examined gene sets associated with high-risk tumors in the RPM using the GSEA (Fig. 3). Pathway enrichment analysis revealed that five significant pathways were enriched for high-risk tumors; cell cycle related gene sets (mitotic spindle; NES = 1.98, FDR q < 0.0001, G2/M check point; NES = 1.83, FDR q = 0.003, E2F targets; NES = 1.76, FDR q = 0.006), MYC target v2 (NES = 1.68, FDR q = 0.006), and PI3K-AKT-mTOR signaling (NES = 1.68, FDR q = 0.012). No gene set was en-

Table 3. Patients and Clinical Characteristics Associated With Recurrence Prediction Model in METABRIC HR$^+$ HER2$^-$ Cohort

Variables	Number of patients (%)			P-value
	Total (N = 1,355)	Recurrence prediction model		
		High risk (N = 486)	Low risk (N = 869)	
Age				
≥ 50	220 (16.2)	58 (11.9)	162 (18.6)	0.0013*
< 50	1,135 (83.8)	428 (88.1)	707 (81.4)	
Menopausal state				
Pre	220 (16.2)	58 (11.9)	162 (18.6)	0.0013*
Post	1,135 (83.8)	428 (88.1)	707 (81.4)	
Tumor size (cm)				
≥ 2	601 (44.4)	179 (36.8)	422 (48.6)	0.000027*
< 2	742 (54.8)	303 (62.3)	439 (50.5)	
Unknown	12 (0.9)	4 (0.8)	8 (0.9)	
Lymph node metastases				
Negative	745 (55)	246 (50.6)	499 (57.4)	0.016*
Positive	610 (45)	240 (49.4)	370 (42.6)	
Histopathology				
Ductal	1,006 (74.2)	395 (81.3)	611 (70.3)	0.000051*
Lobular	118 (8.7)	29 (6)	89 (10.2)	
Others/unknown	231 (17)	62 (12.8)	169 (19.4)	
Tumor grade				
1	159 (11.7)	18 (3.7)	141 (16.2)	< 0.00001*
2, 3	1,135 (83.8)	452 (93)	683 (78.6)	
Unknown	61 (4.5)	16 (3.3)	45 (5.2)	
Clinical stage				
I/II	933 (68.9)	317 (65.2)	616 (70.9)	0.13
III/IV	70 (5.2)	30 (6.2)	40 (4.6)	
Unknown	352 (26)	139 (28.6)	213 (24.5)	
PgR				
Negative	411 (30.3)	206 (42.4)	205 (23.6)	< 0.00001*
Positive	944 (69.7)	280 (57.6)	664 (76.4)	
Molecular characterization				
Luminal A	656 (48.4)	137 (28.2)	519 (59.7)	< 0.00001*
Luminal B	419 (30.9)	222 (45.7)	197 (22.7)	
HER2	63 (4.6)	53 (10.9)	10 (1.2)	
Basal-like	25 (1.8)	22 (4.5)	3 (0.3)	
Claudin-low	72 (5.3)	19 (3.9)	53 (6.1)	
Normal	114 (8.4)	30 (6.2)	84 (9.7)	

*It was also significance in univariate and multivariate analysis. $P < 0.05$ is considered statistically significant. METABRIC: Molecular Taxonomy of Breast Cancer International Consortium, HR$^+$: hormone receptor positive; HER2: human epidermal growth factor receptor 2; PgR: progesterone receptor.

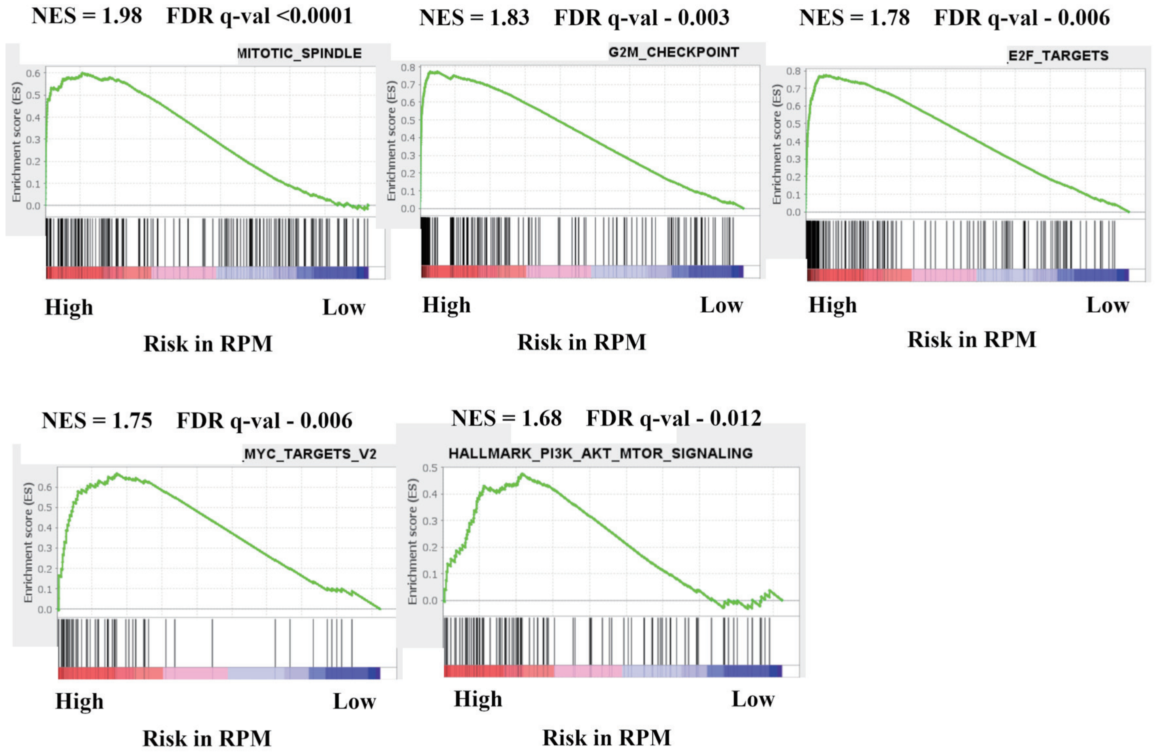

Figure 3. Gene expression profiles based on high and low risk in recurrence prediction model. GSEA of BC patients in METABRIC HR$^+$ HER2$^-$ cohort comparing high and low risk in recurrence prediction model are shown. Upregulated pathways included mitotic spindle, G2/M check point, E2F targets, MYC target v2, and PI3K-AKT-mTOR signaling in high risk compared with low risk in LRM. The significance of each pathway was classified by a threshold of NES > 1.6 or < -1.6 and FDR q-value < 0.025. GSEA: Gene Set Enrichment Analysis; BC: breast cancer; METABRIC: Molecular Taxonomy of Breast Cancer International Consortium; HR$^+$: hormone receptor positive; HER2: human epidermal growth receptor 2; mTOR: mammalian target of rapamycin; LRM: logistic regression model; NES: normalized enrichment score; FDR: false discovery rate.

riched among 50 hallmark gene sets with low-risk tumors.

High-risk tumors were associated with high levels of immune cells from the lymphoid and myeloid lineage and immune cytolytic activity, and low levels of stem cells and stromal cells

We explored the difference in TME composition among high- and low-risk tumors in the RPM utilizing xCell (Fig. 4). High-risk tumors had a higher fraction of immune cells, and a decreased fraction of stem cells and stromal cells. For innate immune cells, eosinophils and macrophages were more prevalent in high-risk tumors, whereas mast cells and monocytes were lower. There was no significant difference in dendritic cells between high- and low-risk tumors in RPM. However, immature dendritic cells and conventional dendritic cells were significantly lower, whereas activated dendritic cells were significantly higher in the high-risk group. CD4$^+$ naive T cells, CD4$^+$ T cells, central memory CD4$^+$ T cells, effector memory CD8$^+$ T cells, class-switched memory B cells, naive B cells, natural killer T cells, pro B cells, Tgd, Th1, Th2 cells, and regulatory T cells (Tregs) were also significantly higher in the high-risk tumor group, but effector CD4$^+$ T cells, central memory CD8$^+$ T cells, and plasma cells were found to be significantly lower. The immune cytolytic activity score has been well established measure of the overall cytolytic activity of immune effector cells in bulk tumors [49]. We found that the immune cytolytic activity scores in BC tumors were significantly higher in the high-risk group. Additionally, in stromal cells, we found that representative stromal cells, such as endothelial cells, were significantly lower amongst the high-risk group. The xCell package enabled the generation of the stroma score using the sums of fractions of certain cell types [48]. We found that stroma scores in BC tumors were significantly lower in the high-risk group. These results indicate that high-risk patients in the RPM had a higher fraction and activity of immune cells, whilst simultaneously having a lower fraction of stem cells and stromal cells in breast TME.

High-risk tumors among patients in the RPM were associated with poor treatment efficacy, which was supported by their correlations with pathways and tumor immune microenvironment (TIME) involved in treatment resistance

Here, we examined the relationship between risk classifica-

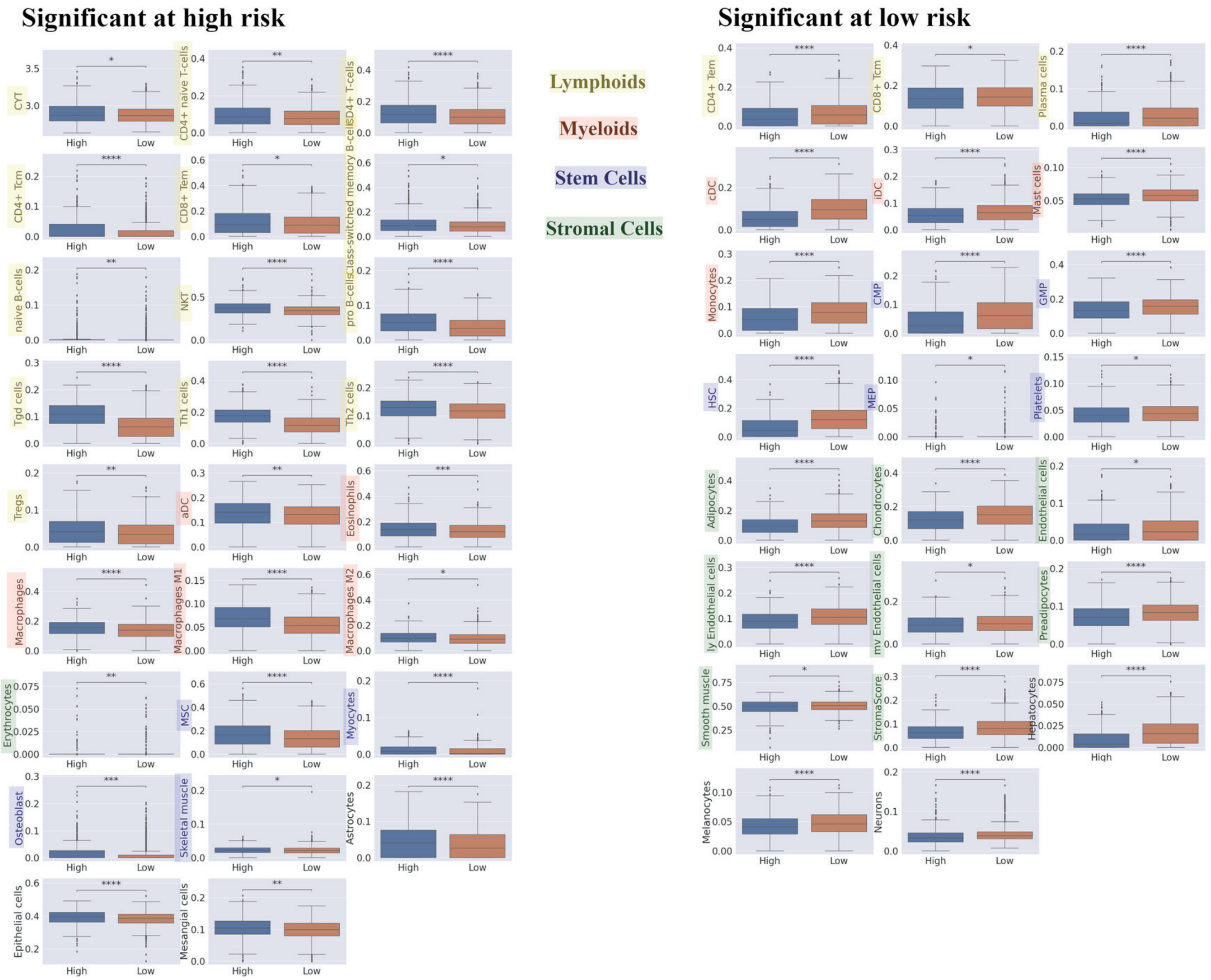

Figure 4. Differences in TME compositions for high and low risk in recurrence prediction model. We explored the difference in TME composition between high- and low- risk in recurrence prediction model utilizing xCell. Box plot of the relationship between recurrence risk in recurrence prediction model and TME in METABRIC HR$^+$ HER2$^-$ cohort are shown. The left panel shows the cell fraction with up-regulation in high risk, and the right panel shows the cell fraction with up-regulation in low risk. ****P < 0.0001, ***P < 0.001, **P < 0.01, *P < 0.05. TME: tumor microenvironment; METABRIC: Molecular Taxonomy of Breast Cancer International Consortium; HR$^+$: hormone receptor positive; HER2: human epidermal growth receptor 2; CYT: immune cytolytic activity; CD4$^+$ tcm: the central memory CD4$^+$ T cell; CD8$^+$ tem: the effector memory CD8$^+$ T cell; NKT: natural killer T cells; Tregs: regulatory T cells; aDC: activated dendritic cell; MSC: mesenchymal stem cell; CD4$^+$ tem: the effector memory CD4$^+$ T cell; CD8$^+$ tcm: the central memory CD8$^+$ T cell; cDC: conventional dendritic cell; iDC: immature dendritic cell; CMP: common myeloid progenitor; GMP: granulocyte-macrophage progenitor; HSC: hematopoietic stem cell; MEP: megakaryocyte-erythroid progenitor.

tion by RPM and treatment prognosis for each recurrence type. The treatment options were ET and CT, which were originally noted in the METABRIC HR$^+$ HER2$^-$ dataset. In distant recurrence and total recurrence analysis, patients in high-risk group were associated with poor prognosis in all treatment groups (Fig. 5). However, in local recurrence analysis, patients in high-risk group were associated with poor prognosis only in the ET group. These results suggest that risk classification by RPM was associated with the effect of treatment, especially the effect of ET in HR$^+$ HER2$^-$ BC patients.

Further, we explored the relationship between treatment outcomes and signaling pathways using GSVA, immune cytolytic activity, and immune cell composition in the METABRIC HR$^+$ HER2$^-$ cohort (Fig. 6). Based on treatments and recurrence, we classified patients into following three groups: patients treated with CT but relapsed as “CT rec”, patients treated with ET alone but relapsed as “ET rec”, and patients treated with ET with or without CT and who did not relapse as “No

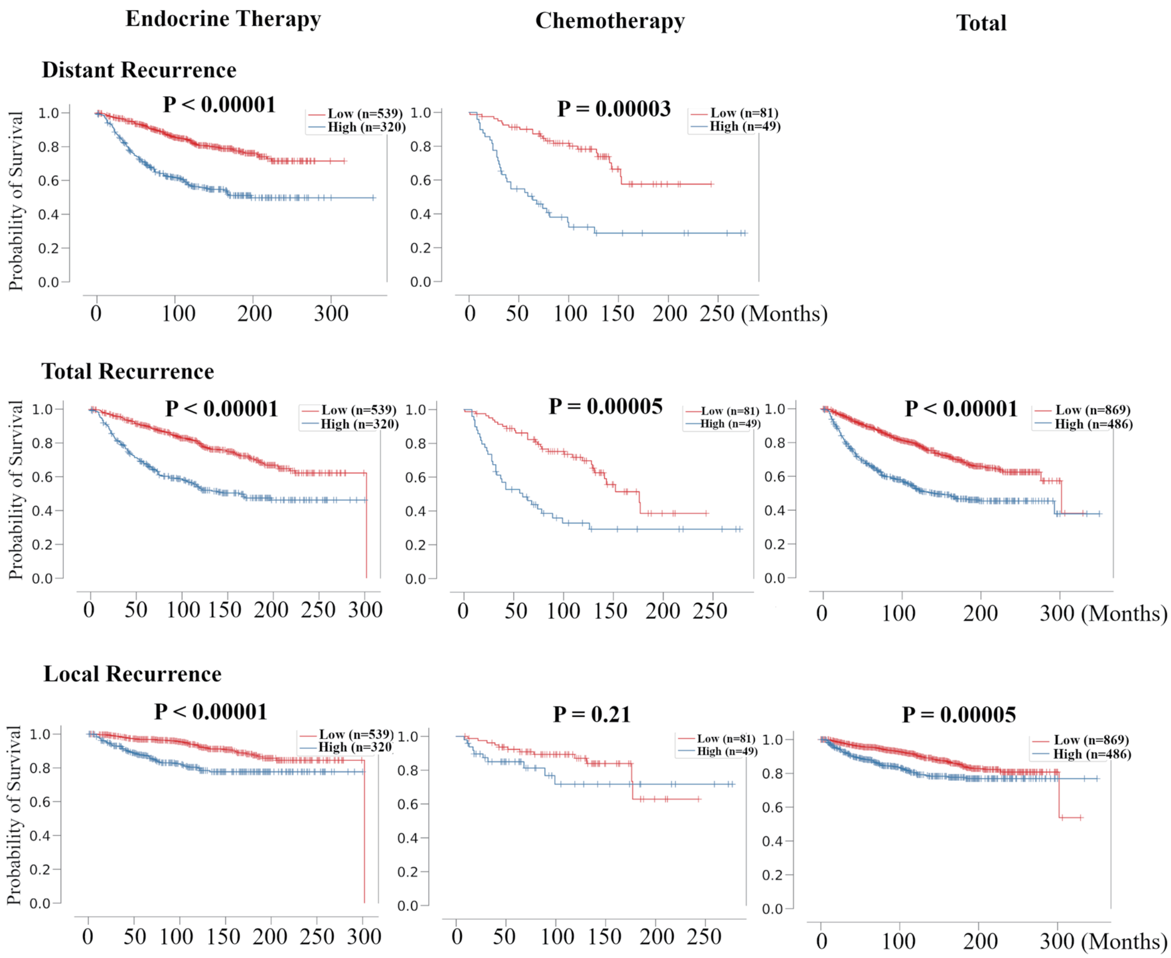

Figure 5. Validation of the relationship between risk classification in recurrence prediction model and the therapeutic effect of chemotherapy and endocrine therapy for HR$^+$ HER2$^-$ BC patients. Kaplan-Meier plots of distant RFS, total RFS, and local RFS of the association between recurrence risk in recurrence prediction model and chemotherapy- and endocrine-treated patients in METABRIC HR$^+$ HER2$^-$ cohort are shown. For total RFS and local RFS, Kaplan-Meier plots of total recurrence are shown on the far right. HR$^+$: hormone receptor positive; HER2: human epidermal growth receptor 2; RFS: recurrence free survival; METABRIC: Molecular Taxonomy of Breast Cancer International Consortium.

rec”. E2F targets, G2M checkpoint, Myc targets v2, and PI3K-AKT-mTOR signaling GSVA scores were highest in the “CT rec” group and lowest in the “No rec” group. High risk of RPM tumors was indeed correlated with these signaling pathways (Fig. 3). Other notable findings were that fatty acid metabolism, protein secretion, and xenobiotic metabolism scores were the highest in the “CT rec” group and the lowest in “ET rec” group. ESTROGEN_RESPONSE_EARLY score was highest in the “No rec” group. In analyzing the relationship amongst the three groups in breast TIME, the immune cytolytic activity scores were the highest in the “CT rec” group. High risk of an RPM tumor was indeed correlated with high immune cytolytic activity scores (Fig. 4). Regarding the immune cell composition, immunostimulatory cells, M1 macrophages were higher in the “CT rec” group compared with the “No rec” group, and follicular helper T cells were lower in the “CT rec” group compared with the “ET rec” group. Immunosuppressive Tregs were the lowest in the “CT rec” group. Additionally, monocytes were found to be higher in the “CT rec” group compared with the “ET rec” group. These results indicate that high-risk tumors among patients in the RPM were associated with poor treatment efficacy, which was supported by their correlation with pathway and TIME, which were involved in treatment resistance.

Discussion

Given the high prevalence and long latency of HR$^+$ BC, the ability to predict prognosis is vital to selecting the optimal therapy for each patient and avoiding overtreatment [4]. Methods to better stratify individuals at high risk for BC development have been a focus of research interest for over a decade [2]. However, none of the tests developed to date are adequate predictors of survival. Therefore, we aimed to test the hypothesis that artificial intelligence can be used to develop a new

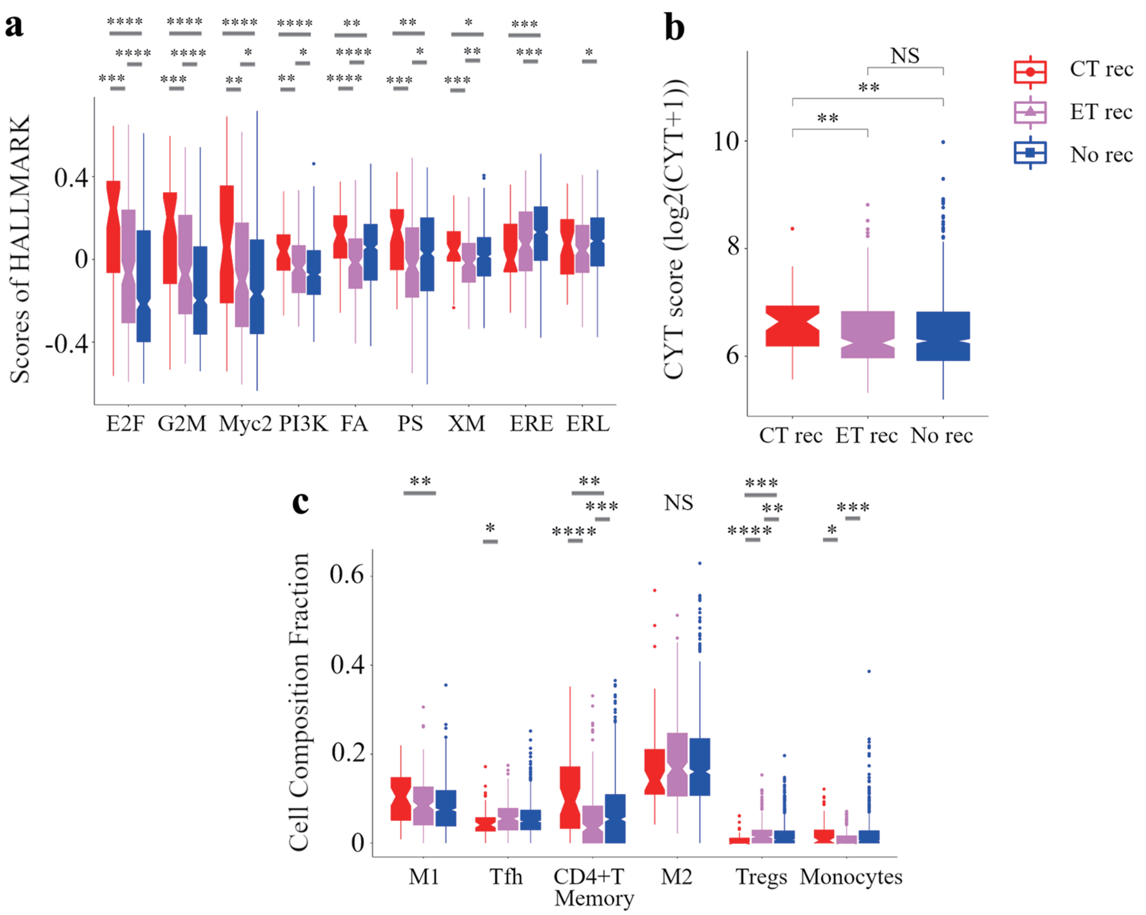

Figure 6. Analysis of the tumor microenvironment with no additional effect of chemotherapy and endocrine therapy. Box plots of the relationship between recurrence by treatments and signaling pathways in GSVA (a) and CYT (b) and immune cell composition (c) in METABRIC HR$^+$ HER2$^-$ cohort are shown. Based on treatments and recurrence, we classified patients into following three categories: a group of patients who were treated with ET and CT but relapsed as CT rec, a group of patients who were treated with ET alone but relapsed as ET rec, and a group of patients who were treated with ET with or without CT and not relapsed as No rec. ****P < 0.0001, ***P < 0.001, **P < 0.01, *P < 0.05. METABRIC: Molecular Taxonomy of Breast Cancer International Consortium; GSVA: Gene Set Variant Analysis; CYT: immune cytolytic activity; ET: endocrine therapy; CT: chemotherapy; E2F: E2F_TARGETS; G2M: G2M_CHECKPOINT; Myc2: MYC_TARGETS_v2; PI3K: PI3K_AKT_MTOR_SIGNALING; FA, FATTY_ACID_METABOLISM; PS: PROTEIN_SECRETION; XM: XENOBIOTIC_METABOLISM; ERE: ESTROGEN_RESPONSE_EARLY; ERL: ESTROGEN_RESPONSE_LATE; WNTβ: WNT_BETA_CATENIN_SIGNALING; M1: M1 macrophage; Tfh: follicular helper cells; M2: M2 macrophage; Tregs: CD4$^+$ regulatory T cells.

framework for predicting prognosis in BC patients in multiple validation BC cohorts.

This study generated three interesting results with clinical implications. First, our RPM was able to stratify patients by prognosis in multiple cohorts. We first extracted 155 prognostic-related genes through a meta-analysis of all protein-coding genes associated with distant RFS in the METABRIC HR$^+$ HER2$^-$ cohort (Supplementary Material 2, www.wjon.org). we combined the Cox-PH model with recursive feature elimination to narrow down the nine best predictors (Table 2). We next applied an LRM to develop tools that can accurately predict recurrence risk in BC patients based on the binary expression status of these nine genes.

Interestingly, most of these nine genes have not been well characterized to date with regard to their relation to basic or clinical oncology, with *CYP4F22*, *TUBA3D*, and *HNMT* apparently not having been studied in the field at all. Therefore, further research will be needed to elucidate the effects of nine genes based on RPM to extend the prognosis of BC patients. Finally, we found that our RPM can universally and predominantly stratify prognosis in all HR$^+$ HER2$^-$ BC cohorts, using the Kaplan-Meier method and verified by the log-rank test (Fig. 2). In a similar study, Shimizu and Nakayama reported generating a prognostic score with a neural network for 23 genes narrowed down by one of the machine learnings, random forest, from about 20,000 genes [52]. The combination of the score and clinical stage suggested that prognosis could be stratified more precisely and that unnecessary CT could be avoided.

Further, the time-dependent AUC analysis of RPM showed a trend towards a significant decrease in accuracy around 10 years, although there was some variability between cohorts (Supplementary Material 4, www.wjon.org). In general, the HR$^+$ BC subtype has a longer recurrence period than the HER2-enriched or triple-negative subtypes. Thus, prediction of late recurrence is one of the most important factors in predicting recurrence of HR$^+$ BC [5, 16]. Therefore, to further improve the accuracy of our model, it is necessary to create a model that considers not only early recurrence, but late recurrence as well.

Second, the high-risk group in the RPM was associated with major BC therapeutic target pathways and TIME. We have shown that cell cycle-associated gene sets, MYC targets, and PI3K-AKT-mTOR signaling in GSEA were enriched in the high-risk group in the RPM (Fig. 3). We and many other research groups have previously reported that genetic mutations that cause abnormalities in the PI3K-AKT-mTOR pathway are closely related to the prognosis of BC patients [53-58]. Further, recent advances in sequencing of the human genome have revealed cell cycle and PI3K-AKT-mTOR pathway as therapeutic targets for HR$^+$ HER2$^-$ metastatic BC, ranging from hormone therapy single agents to CDK4/6 inhibitors (palbociclib, ribociclib, abemaciclib), and an mTOR inhibitor (everolimus) [59, 60]. The cell cycle has also been identified as an important therapeutic target in primary BC as the monarchE trial demonstrated that abemaciclib was particularly useful as an adjuvant therapy [61].

In the component analysis of TME, the high-risk group in the RPM was associated with high levels of immune cells from the lymphoid and myeloid lineage and immune cytolytic activity, as well as low levels of stem and stromal cells (Fig. 4). Tumor-infiltrating lymphocytes are broadly classified into CD4$^+$ helper cells, Tregs, and effector cells such as natural killer cells and CD8$^+$ T cells [62]. Tregs, which normally suppress autoreactive T cells, suppress anti-tumor responses in TME and are therefore a poor prognostic factor in BC [63, 64]. Conversely, infiltration of CD8$^+$ effector T cells into tumors is associated with longer BC-specific survival, independent of other prognostic factors such as tumor grade, clinical stage, and vascular invasion [63, 65]. Regarding the immune cytolytic activity score, which indicates the relative degree of anticancer immune activity, we have reported that it functions as a prognostic biomarker for BC [16, 23, 24]. Surprisingly, the high-risk group in the RPM had a mixture of good and poor prognostic factors, as shown by high correlations with immune cytolytic activity, Th1, Th2, Tregs, and the effector memory CD8$^+$ T cells.

On the other hand, regarding myeloid-derived immune cells, macrophages are mostly of the M1 phenotype during normal immune response and are involved in the Th1 cytokine response to various pathogens. However, the tumor-associated macrophages that are formed in breast tumors and typically belong to the M2 phenotype, allow for cancer cell survival, and have been positioned as a poor prognostic factor in BC [62, 63]. Dendritic cells have different roles in tumors depending on their degree of maturation. Tumor-associated immature dendritic cells produce pro-angiogenic factors and actively promote tumor growth. Mature dendritic cells activate CD4$^+$ and CD8$^+$ T cells to attack tumor cells and reduce metastasis [62, 66]. Similar to lymphoids, the high-risk group in the RPM had a mixture of good and poor prognostic factors, as shown by high correlations with activated dendritic cells, M1 macrophages, and M2 macrophages. A possible reason for the discrepancy between the high-risk group in the RPM and immune cell function, is that immune response in tumor tissue is a series of carefully controlled events that can be optimally addressed as a group rather than as individual cells [67].

Third, the high-risk group in the RPM was associated with poor treatment efficacy, which was supported by their correlations with pathways and TIME involved in treatment resistance. We demonstrate that risk classification by RPM was associated with the effect of treatment, especially the effect of ET in HR$^+$ HER2$^-$ BC patients (Fig. 5). In the analysis of the relationship between treatment outcome and signaling pathways, immune cytolytic activity, and immune cell composition, HR$^+$ HER2$^-$ BC patients treated with CT but who relapsed had higher levels of E2F targets and G2M checkpoints, and were associated with lower levels of immunosuppressive M2 macrophages and Tregs (Fig. 6). According to a widely accepted idea, both cancer cell-specific properties, as well as signals derived from cells in TME play a critical role in cancer therapy response [68]. For example, Rosenfeldt and his colleagues showed that the E2F1-ABCG2 axis suppresses the CT-induced cell death that can be restored by the inhibition of ABCG2 [69]. Further, Velaei and colleagues reviewed the impact of conventional anticancer CT on the relationship between the tumor and the immune system as follows [68]. High levels of the tumor-infiltrating lymphocytes in the HR$^-$ HER2$^+$ subtype are related with good response to CT. Above all, high levels of CD8$^+$ cytotoxic T lymphocytes synergistically increase the effect of anthracycline or anthracycline/taxane based neoadjuvant CT and low levels of Tregs enhance CT response. In addition, targeting tumor-associated macrophages in combination with CT may improve the effect of CT, since macrophages induced by cytotoxic CT can protect tumor cells from death due to a cathepsin-dependent function. Focusing on the importance of gene expression profiles in the tumor stroma of BC patients, Finak et al generated a 26-gene prognostic predictor that predicts clinical outcomes regardless of clinical subtype [70]. Their model was correlated with two different sets of relevant genes: hypoxia and angiogenesis. These were associated with poor prognosis or displayed a Th1-like immune response associated with favorable outcomes. We demonstrate that high-risk status in the RPM was associated with poor treatment efficacy, which was supported by their correlation with E2F targets, G2M checkpoints, MYC targets, and PI3K-AKT-mTOR signaling, immune cytolytic activity, M1 macrophages, and Tregs, which were involved in treatment resistance. Further research is needed to create models that stratify prognosis by therapeutic factor in estrogen receptor (ER)$^+$ HER2$^-$ BC and to identify novel strategies to overcome therapeutic resistance.

Although the study demonstrates promising results, it has limitations. First, this study utilized multiple large patient cohorts, and is therefore a retrospective study. Secondly, the investigated cohorts were not genetically analyzed using a common platform. Finally, the cohorts other than METABRIC lacked data on treatment regimens and the relationship between the RPM and treatment outcomes could not be investigated.

In conclusion, using machine learning, we identified nine genes related to BC prognosis from more than 18,000 genes and created an RPM. The RPM was able to stratify prognosis in multiple cohorts. This is because the RPM reflects major BC therapeutic target pathways and TIME and, further supported by the therapeutic effect of CT and ET in patients with BC. Based on these reported results, we anticipate that further studies can be conducted to better understand the mechanisms of recurrence and resistance to therapy in BC.

Supplementary Material

Suppl 1. Patients and clinical characteristics in TCGA HR$^+$ HER2$^-$ cohort and other independent HR$^+$ HER2$^-$ BC cohorts, GSE199135, GSE9195, GSE6532, and GSE21653. TCGA: The Cancer Genome Atlas; HR$^+$: hormone receptor positive; HER2: human epidermal growth receptor 2; BC: breast cancer; PgR: progesterone receptor.

Suppl 2. 155 differentially expressed genes with |log2-fold change| > 0.25 and P < 0.05 in distant recurrence of HR$^+$ HER2$^-$ breast cancer. HR$^+$: hormone receptor positive; HER2: human epidermal growth receptor 2.

Suppl 3. 23 genes closely involved in distant RFS in HR$^+$ HER2$^-$ patients extracted from 155 signature genes using the Cox-PH model. RFS: recurrence free survival; HR$^+$: hormone receptor positive; HER2: human epidermal growth receptor 2; Cox-PH: Cox Proportional-Hazards.

Suppl 4. Analysis of changes in predictive accuracy of recurrence prediction model over clinical course. Time-dependent receiver operating characteristic curve depicting the time-dependent AUC values of the recurrence prediction model in METABRIC HR$^+$ HER2$^-$ cohort, TCGA HR$^+$ HER2$^-$ cohort, GSE199135, GSE9195, GSE6532, and GSE21653 are shown. AUC: area under curve; METABRIC: Molecular Taxonomy of Breast Cancer International Consortium; HR$^+$: hormone receptor positive; HER2: human epidermal growth receptor 2; TCGA: The Cancer Genome Atlas.

Acknowledgments

None to declare.

Financial Disclosure

This work was supported by NIH grant R01CA160688 to KT, and NCI grant P30CA016056 involving the use of Roswell Park Cancer Institute's Bioinformatics and Biostatistics Shared Resources.

Conflict of Interest

All of the authors declare that they have no actual, potential, or perceived conflict of interest regarding the manuscript submitted for review.

Informed Consent

Not applicable.

Author Contributions

Takashi Takeshita: conceptualization, methodology, doftware, writing - original draft preparation, visualization. Hirotaka Iwase: Writing - reviewing and editing. Rongrong Wu: software, validation. Danya Ziazadeh: writing - reviewing and editing. Li Yan: data curation, software, validation. Kazuaki Takabe: supervision, funding acquisition.

Data Availability

The data presented in this study are available upon request from the corresponding author.

Abbreviations

BC: breast cancer; HR$^+$: hormone receptor positive; HER2: human epidermal growth factor receptor 2; ET: endocrine therapy; CT: chemotherapy; METABRIC: Molecular Taxonomy of Breast Cancer International Consortium; TCGA: The Cancer Genome Atlas; GEO: Gene Expression Omnibus; Cox-PH: Cox proportional-hazards; RFS: recurrence free survival; LRM: logistic regression model; NES: normalized enrichment score; FDR: false discovery rate; GSVA: Gene Set Variation Analysis; TME: tumor microenvironment; DEGs: differentially expressed genes; AUC: area under the receiver operating characteristic curve; PgR: progesterone receptor; GSEA: Gene Set Enrichment Analysis; mTOR: mammalian target of rapamycin; Tregs: regulatory T cells; TIME: tumor immune microenvironment

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72(1):7-33. doi pubmed
2. Esteva FJ, Sahin AA, Cristofanilli M, Arun B, Hortobagyi GN. Molecular prognostic factors for breast cancer metastasis and survival. Semin Radiat Oncol. 2002;12(4):319-328. doi pubmed
3. Hayes DF. Clinical utility of genetic signatures in selecting adjuvant treatment: Risk stratification for early vs. late recurrences. Breast. 2015;24(Suppl 2):S6-S10. doi pubmed
4. Kwa M, Makris A, Esteva FJ. Clinical utility of gene-expression signatures in early stage breast cancer. Nat Rev Clin Oncol. 2017;14(10):595-610. doi pubmed
5. Puppe J, Seifert T, Eichler C, Pilch H, Mallmann P, Malter W. Genomic signatures in luminal breast cancer. Breast Care (Basel). 2020;15(4):355-365. doi pubmed pmc

6. Zhang W, Zhao Z, Wang K, Shen L, Shi X. The International Conference on Intelligent Biology and Medicine (ICIBM) 2020: Scalable techniques and algorithms for computational genomics. BMC Genomics. 2020;21(Suppl 11):831. doi pubmed pmc
7. Munir A, Vedithi SC, Chaplin AK, Blundell TL. Genomics, computational biology and drug discovery for mycobacterial infections: fighting the emergence of resistance. Front Genet. 2020;11:965. doi pubmed pmc
8. Becker M, Schultze H, Bresniker K, Singhal S, Ulas T, Schultze JL. A novel computational architecture for large-scale genomics. Nat Biotechnol. 2020;38(11):1239-1241. doi pubmed
9. Krop I, Ismaila N, Andre F, Bast RC, Barlow W, Collyar DE, Hammond ME, et al. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American Society of Clinical Oncology Clinical Practice Guideline Focused Update. J Clin Oncol. 2017;35(24):2838-2847. doi pubmed pmc
10. Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature. 2012;486(7403):405-409. doi pubmed pmc
11. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell. 2015;163(2):506-519. doi pubmed pmc
12. Gupta A, Mutebi M, Bardia A. Gene-expression-based predictors for breast cancer. Ann Surg Oncol. 2015;22(11):3418-3432. doi pubmed
13. Kamel HFM, Al-Amodi H. Exploitation of gene expression and cancer biomarkers in paving the path to era of personalized medicine. Genomics Proteomics Bioinformatics. 2017;15(4):220-235. doi pubmed pmc
14. Katsuta E, Yan L, Nagahashi M, Raza A, Sturgill JL, Lyon DE, Rashid OM, et al. Doxorubicin effect is enhanced by sphingosine-1-phosphate signaling antagonist in breast cancer. J Surg Res. 2017;219:202-213. doi pubmed pmc
15. Young J, Kawaguchi T, Yan L, Qi Q, Liu S, Takabe K. Tamoxifen sensitivity-related microRNA-342 is a useful biomarker for breast cancer survival. Oncotarget. 2017;8(59):99978-99989. doi pubmed pmc
16. Takeshita T, Yan L, Asaoka M, Rashid O, Takabe K. Late recurrence of breast cancer is associated with pro-cancerous immune microenvironment in the primary tumor. Sci Rep. 2019;9(1):16942. doi pubmed pmc
17. Takeshita T, Asaoka M, Katsuta E, Photiadis SJ, Narayanan S, Yan L, Takabe K. High expression of polo-like kinase 1 is associated with TP53 inactivation, DNA repair deficiency, and worse prognosis in ER positive Her2 negative breast cancer. Am J Transl Res. 2019;11(10):6507-6521. pubmed pmc
18. Asaoka M, Patnaik SK, Zhang F, Ishikawa T, Takabe K. Lymphovascular invasion in breast cancer is associated with gene expression signatures of cell proliferation but not lymphangiogenesis or immune response. Breast Cancer Res Treat. 2020;181(2):309-322. doi pubmed pmc
19. Tokumaru Y, Oshi M, Katsuta E, Yan L, Satyananda V, Matsuhashi N, Futamura M, et al. KRAS signaling enriched triple negative breast cancer is associated with favorable tumor immune microenvironment and better survival. Am J Cancer Res. 2020;10(3):897-907. pubmed pmc
20. Asaoka M, Ishikawa T, Takabe K, Patnaik SK. APOBEC3-Mediated RNA editing in breast cancer is associated with heightened immune activity and improved survival. Int J Mol Sci. 2019;20(22):5621. doi pubmed pmc
21. Katsuta E, Maawy AA, Yan L, Takabe K. High expression of bone morphogenetic protein (BMP) 6 and BMP7 are associated with higher immune cell infiltration and better survival in estrogen receptor-positive breast cancer. Oncol Rep. 2019;42(4):1413-1421. doi pubmed pmc
22. McDonald KA, Kawaguchi T, Qi Q, Peng X, Asaoka M, Young J, Opyrchal M, et al. Tumor heterogeneity correlates with less immune response and worse survival in breast cancer patients. Ann Surg Oncol. 2019;26(7):2191-2199. doi pubmed pmc
23. Takeshita T, Torigoe T, Yan L, Huang JL, Yamashita H, Takabe K. The impact of immunofunctional phenotyping on the malfunction of the cancer immunity cycle in breast cancer. Cancers (Basel). 2020;13(1):110. doi pubmed pmc
24. Takeshita T, Tokumaru Y, Oshi M, Wu R, Patel A, Tian W, Hatanaka Y, et al. Clinical relevance of estrogen reactivity in the breast cancer microenvironment. Front Oncol. 2022;12:865024. doi pubmed pmc
25. Osipowski P, Pawelkowicz M, Wojcieszek M, Skarzynska A, Przybecki Z, Plader W. A high-quality cucumber genome assembly enhances computational comparative genomics. Mol Genet Genomics. 2020;295(1):177-193. doi pubmed
26. Alberghina L, Piccialli G. From computational genomics to systems metabolomics for precision cancer medicine and drug discovery. Pharmacol Res. 2020;151:104479. doi pubmed
27. Lacorte GA, Machado MA, Martinez ML, Campos AL, Maciel RP, Verneque RS, Teodoro RL, et al. DGAT1 K232A polymorphism in Brazilian cattle breeds. Genet Mol Res. 2006;5(3):475-482. pubmed
28. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261-272. doi pubmed pmc
29. Hayward P. Word recognition localised to left occipitotemporal cortex. Lancet Neurol. 2006;5(6):475. doi pubmed
30. Renfree SP, Makovicka JL, Chung AS. Risk factors for delay in surgery for patients undergoing elective anterior cervical discectomy and fusion. J Spine Surg. 2019;5(4):475-482. doi pubmed pmc
31. http://www.cbioportal.org/.
32. https://www.ncbi.nlm.nih.gov/geo/.
33. https://www.python.org.
34. https://numpy.org.
35. https://scipy.org.
36. https://pandas.pydata.org.

37. https://seaborn.pydata.org.
38. https://pypi.org/project/matplotlib/.
39. https://www.r-project.org.
40. Oshi M, Takahashi H, Tokumaru Y, Yan L, Rashid OM, Nagahashi M, Matsuyama R, et al. The E2F pathway score as a predictive biomarker of response to neoadjuvant therapy in ER+/HER2- breast cancer. Cells. 2020;9(7):1643. doi pubmed pmc
41. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401-404. doi pubmed pmc
42. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1. doi pubmed pmc
43. Rueda OM, Sammut SJ, Seoane JA, Chin SF, Caswell-Jin JL, Callari M, Batra R, et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. Nature. 2019;567(7748):399-404. doi pubmed pmc
44. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell. 2018;173(2):400-416.e411. doi pubmed pmc
45. Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. Journal of Mathematical Modelling and Algorithms. 2006;5:475-504.
46. http://software.broadinstitute.org/gsea/index.jsp.
47. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417-425. doi pubmed pmc
48. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol. 2017;18(1):220. doi pubmed pmc
49. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015;160(1-2):48-61. doi pubmed pmc
50. Balli D, Rech AJ, Stanger BZ, Vonderheide RH. Immune cytolytic activity stratifies molecular subsets of human pancreatic cancer. Clin Cancer Res. 2017;23(12):3129-3138. doi pubmed
51. http://bioconductor.org/.
52. Shimizu H, Nakayama KI. A 23 gene-based molecular prognostic score precisely predicts overall survival of breast cancer patients. EBioMedicine. 2019;46:150-159. doi pubmed pmc
53. Hortobagyi GN, Chen D, Piccart M, Rugo HS, Burris HA, 3rd, Pritchard KI, Campone M, et al. Correlative analysis of genetic alterations and everolimus benefit in hormone receptor-positive, human epidermal growth factor receptor 2-negative advanced breast cancer: results from BOLERO-2. J Clin Oncol. 2016;34(5):419-426. doi pubmed pmc
54. Takeshita T, Yamamoto Y, Yamamoto-Ibusuki M, Tomiguchi M, Sueta A, Murakami K, Iwase H. Clinical significance of plasma cell-free DNA mutations in PIK3CA, AKT1, and ESR1 gene according to treatment lines in ER-positive breast cancer. Mol Cancer. 2018;17(1):67. doi pubmed pmc
55. Takeshita T, Yamamoto Y, Yamamoto-Ibusuki M, Tomiguchi M, Sueta A, Iwase H. ESR1 and PIK3CA mutational status in serum and plasma from metastatic breast cancer patients: A comparative study. Cancer Biomark. 2018;22(2):345-350. doi pubmed
56. Takeshita T, Yamamoto Y, Yamamoto-Ibusuki M, Tomiguchi M, Sueta A, Murakami K, Omoto Y, et al. Comparison of ESR1 mutations in tumor tissue and matched plasma samples from metastatic breast cancer patients. Transl Oncol. 2017;10(5):766-771. doi pubmed pmc
57. Takeshita T, Yamamoto Y, Yamamoto-Ibusuki M, Tomiguchi M, Sueta A, Murakami K, Omoto Y, et al. Analysis of ESR1 and PIK3CA mutations in plasma cell-free DNA from ER-positive breast cancer patients. Oncotarget. 2017;8(32):52142-52155. doi pubmed pmc
58. Takeshita T, Yamamoto Y, Yamamoto-Ibusuki M, Inao T, Sueta A, Fujiwara S, Omoto Y, et al. Prognostic role of PIK3CA mutations of cell-free DNA in early-stage triple negative breast cancer. Cancer Sci. 2015;106(11):1582-1589. doi pubmed pmc
59. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921. doi pubmed
60. McAndrew NP, Finn RS. Clinical review on the management of hormone receptor-positive metastatic breast cancer. JCO Oncol Pract. 2022;18(5):319-327. doi pubmed
61. Johnston SRD, Harbeck N, Hegg R, Toi M, Martin M, Shao ZM, Zhang QY, et al. Abemaciclib combined with endocrine therapy for the adjuvant treatment of HR+, HER2-, node-positive, high-risk, early breast cancer (monarchE). J Clin Oncol. 2020;38(34):3987-3998. doi pubmed pmc
62. Zhang AW, Campbell KR. Computational modelling in single-cell cancer genomics: methods and future directions. Phys Biol. 2020;17(6):061001. doi pubmed
63. Grosjean N, Blaby-Haas CE. Leveraging computational genomics to understand the molecular basis of metal homeostasis. New Phytol. 2020;228(5):1472-1489. doi pubmed
64. Caroli J, Dori M, Bicciato S. Computational methods for the integrative analysis of genomics and pharmacological data. Front Oncol. 2020;10:185. doi pubmed pmc
65. Krishnan A, Kloehn J, Lunghi M, Chiappino-Pepe A, Waldman BS, Nicolas D, Varesio E, et al. Functional and computational genomics reveal unprecedented flexibility in stage-specific toxoplasma metabolism. Cell Host Microbe. 2020;27(2):290-306.e211. doi pubmed
66. Zhang C, Mathe E, Ning X, Zhao Z, Wang K, Li L, Guo Y. The International Conference on Intelligent Biology and Medicine 2019 (ICIBM 2019): computational methods and applications in medical genomics. BMC Med Genomics. 2020;13(Suppl 5):47. doi pubmed pmc
67. Chen DS, Mellman I. Oncology meets immunology: the

cancer-immunity cycle. Immunity. 2013;39(1):1-10. doi pubmed

68. Velaei K, Samadi N, Barazvan B, Soleimani Rad J. Tumor microenvironment-mediated chemoresistance in breast cancer. Breast. 2016;30:92-100. doi pubmed

69. Rosenfeldt MT, Bell LA, Long JS, O'Prey J, Nixon C, Roberts F, Dufes C, et al. E2F1 drives chemotherapeutic drug resistance via ABCG2. Oncogene. 2014;33(32):4164-4172. doi pubmed

70. Osipowski P, Pawelkowicz M, Wojcieszek M, Skarzynska A, Przybecki Z, Plader W. Correction to: A high-quality cucumber genome assembly enhances computational comparative genomics. Mol Genet Genomics. 2020;295(2):535. doi pubmed